STATISTICS IN TRANSITION new series, Summer 2015
Vol. 16, No. 2, pp. 163–182

AN APPROXIMATION TO THE OPTIMAL SUBSAMPLE ALLOCATION FOR SMALL AREAS

W. B. Molefe[1], D. K. Shangodoyin[2], R. G. Clark[3]

ABSTRACT

This paper develops allocation methods for stratified sample surveys in which small area estimation is a priority. We assume stratified sampling with small areas as the strata. Similar to Longford (2006), we seek efficient allocation that minimizes a linear combination of the mean squared errors of composite small area estimators and of an estimator of the overall mean. Unlike Longford, we define mean-squared error in a model-assisted framework, allowing a more natural interpretation of results using an intra-class correlation parameter. This allocation has an analytical form for a special case, and has the unappealing property that some strata may be allocated no sample. We derive a Taylor approximation to the stratum sample sizes for small area estimation using composite estimation giving priority to both small area and national estimation.

Key words: composite estimation, mean squared error, sample design, small area estimation, sample size allocation, Taylor approximation.

1. Introduction

Sampling designs, and sample sizes in particular, are chosen in practice so as to provide reliable estimates for large geographical regions or broad demographic groups. Budget and other constraints usually prevent the allocation of sufficiently large samples to each of the small areas. It is not possible to anticipate and plan for all possible areas (or domains) of applications as "the client will always require more than is specified at the design stage" (Fuller, 1999). The increased emphasis on small area estimation raises the question of how best to design surveys when the precision of small area estimates is a priority. If small area data needs are to be served using survey data then there is a need to develop an overall strategy that involves careful attention to satisfy these needs at the planning, sample design and estimation stages of the survey process (Singh et al., 1994). Singh et al. (1994) presented an illustration of compromise sample size allocation

[1] Department of Statistics, University of Botswana. E-mail: molefewb@mopipi.ub.bw.
[2] Department of Statistics, University of Botswana. E-mail: shangodoyink@mopipi.ib.bw.
[3] National Institute for Applied Statistics Research Australia, University of Wollongong. E-mail: rclark@uow.edu.au.

to satisfy reliability requirements at the provincial level as well as sub-provincial level in Canada.

Assume that small areas are identified in advance, and that stratified sampling is used with H strata defined by the small areas, indexed by $h \in U^1$. The population of units, indexed by j, denoted U is of size N. The population of N_h units in stratum h is U_h and the sample of n_h units selected by simple random sampling without replacement (SRSWOR) from stratum h is denoted by s_h. Let Y_j be the value of the characteristic of interest for the j^{th} unit in the population. The small area population mean is $\bar{Y}_h$ and the national mean is $\bar{Y}$. The corresponding sample estimators are $\bar{y}_h$ and $\bar{y}$, respectively; $\bar{y}_h = n_h^{-1} \sum_{j \in s_h} y_j$ and $\bar{y} = \sum_{h \in U^1} P_h \bar{y}_h$, where $P_h = N_h N^{-1}$. Let the sampling variances be $v_h = \text{var}_p(\bar{y}_h)$ and $v = \text{var}_p(\bar{y})$

Longford (2006) considers the problem of optimal sample sizes for small area estimation for this design. The approach is based on minimizing the weighted sum of the mean squared errors of the planned small area mean estimates and an overall estimate of the mean, with the weights specified to reflect the inferential priorities. An analytical solution exists when no weight is attached to estimating the overall mean but it has undesirable practical properties. For example, the optimal sample allocation is arrived at iteratively and some stratum sample sizes may be zero. When the overall mean is also important Longford does not find an exact or approximate analytical solution to the optimization problem. He suggests that the equation can be solved by numerical methods, such as the Newton method which interpolates between or extrapolates from a pair of provisional solutions, but that solving these equations iteratively may involve a considerable amount of computing when there are thousands of small areas.

The aim of this paper is to find the best allocation to strata for a linear combination of the mean squared errors of small area composite estimators and of an overall estimator of the mean, similar to Longford (2006). In section 2 we reformulate the objective in model-assisted terms, and derive the model-assisted composite estimator. Section 3 is devoted to optimizing the design. In subsection 3.1 we derive the optimal allocation for this objective when national estimation has no priority (G = 0) (similar in form to Longford but with different interpretation due to the explicit use of a model). Longford (2006) did not give an analytical solution when both national (overall) and small area estimates are a priority (G > 0). A numerical algorithm was given but may be computationally intensive, and its iterative nature makes it less transparent. In subsection 3.2 we derive two Taylor series approximations to the optimum. Unfortunately, the optimal allocations (both when G = 0 and when G > 0) have some undesirable properties.

2. Composite estimation

Royall (1973), in a discussion of papers by Gonzalez (1973) and Ericksen (1973), suggested that a choice between direct and synthetic approaches need not be made but that '*... a combination of the two is better than either taken alone*'. A natural way to balance the potential bias of a synthetic estimator $\bar{y}$ for $\bar{Y}_h$ against the instability of a direct estimator $\bar{y}_h$, is to use a composite estimator $\bar{y}_h^C$.

Composite estimators for small areas are defined as convex combinations of direct (unbiased) and synthetic (biased) estimators. A simple example is the composition $\bar{y}_h^C = (1-\phi_h)\bar{y}_h + \phi_h \bar{y}$ of the sample mean $\bar{y}_h$ for the target area h and the overall sample mean $\bar{y}$ of the target variable. The (area-specific) coefficients ϕ_h and $1-\phi_h$ in this composition are set with the intent to minimize its mean squared error (MSE), see for example, Schaible (1978); Brock et al. (1980) and Rao (2003). The coefficients for which minimum MSE would be attained depend on some unknown parameters which have to be estimated.

The design-based MSE of the composite estimator is given by:

$$MSE_p\left(\bar{y}_h^C; \bar{Y}_h\right) = (1-\phi_h)^2 v_h + \phi_h^2 \left\{v + B_h^2\right\} + 2\phi_h\left(1-\phi_h\right)C_h \tag{1}$$

where C_h is the sampling covariance of $\bar{y}_h$ and $\bar{y}$, v_h is the sampling variance of the direct estimator $\bar{y}_h$, v is the sampling variance of the synthetic estimator $\bar{y}$ for $\bar{y}_h$ and $B_h = \bar{Y}_h - \bar{Y}$ is the bias of using $\bar{y}$ to estimate $\bar{y}_h$. Further,

$$MSE_p\left(\bar{y}_h^C; \bar{Y}_h\right) \approx (1-\phi_h)^2 v_h + \phi_h^2 B_h^2 \tag{2}$$

because $C_h \ll v_h$ and $v \ll v_h$ when the sample for area h is not a large part of the national sample. Auxiliary variables x_j are assumed to be available for the full population $j \in U$[1].

The following model ξ will be assumed:

$$\left.\begin{aligned} &E_\xi\left[Y_j\right] = \beta^T x_j \\ &\mathrm{var}_\xi\left[Y_j\right] = \sigma^2\left(j \in U_d\right) \\ &\mathrm{cov}_\xi\left[Y_i, Y_j\right] = \rho\sigma^2\left(i \neq j; i, j \in U_d\right) \\ &\mathrm{cov}_\xi\left[Y_i, Y_j\right] = 0\left(i \in U_d, j \in U_k, d \neq k\right) \end{aligned}\right\} \tag{3}$$

where i and j are units and h and g are small areas. Under the model (3),

$$E_\xi\left[v_h\right]=E_\xi\left[\operatorname{var}_p\left(\bar{y}_h\right)\right]=E_\xi\left[n_h^{-1}S_{hw}^2\right]=n_h^{-1}\sigma^2\left(1-\rho\right)$$

and

$$E_\xi\left[B_h^2\right]=E_\xi\left[\left(\bar{Y}_h-\bar{Y}\right)^2\right]\approx \operatorname{var}_\xi\left[\bar{Y}_h\right]=\operatorname{var}_\xi\left(N_h^{-1}\sum\nolimits_{j\in U_h}Y_j\right)=\sigma^2N_h^{-1}\left[1+\left(N_h-1\right)\rho\right]$$

While the areas may have small sample sizes, their population sizes are substantial, so that $E_\xi\left[B_h^2\right]\approx\sigma^2\rho$. Also,

$$E_\xi\left[v\right]=E_\xi\left[\operatorname{var}_p\left(\bar{y}\right)\right]=E_\xi\left(\sum\nolimits_{h\in U^1}P_h^2n_h^{-1}S_{hw}^2\right)=\sigma^2\left(1-\rho\right)\sum\nolimits_{h\in U^1}P_h^2n_h^{-1} \quad (4)$$

Following Molefe and Clark (2015), we assume a small-area composite estimator which is a weighted mean of an approximately design unbiased estimator

$$\bar{y}_{hr}=\bar{y}_h+\beta^T\left(\bar{X}_h-\bar{x}_h\right)$$

recommended by Hidiroglou and Patak (2004) for small domains, and a model-based synthetic estimator $\bar{Y}_{h(syn)}=\beta^T\bar{X}_h$. The composite estimator which approximately minimizes the anticipated MSE is

$$\bar{y}_h^C=(1-\phi_{h(opt)})\bar{y}_{hr}+\phi_{d(opt)}\bar{Y}_{h(syn)}=\beta^T\bar{X}_h+(1-\phi_{h(opt)})\left(\bar{y}_h-\beta^T\bar{x}_h\right)$$

where $\phi_{h(opt)}=(1-\rho)\left[1+(n_h^*-1)\rho\right]^{-1}$, assuming that n, N_h and H are all large (Molefe and Clark, 2015). Under the same assumptions, the approximate anticipated MSE of the optimal composite estimator of $\bar{Y}_h$ conditional on n_h^* is

$$\begin{aligned}&E_\xi MSE_p\left(\bar{y}_h^C\left[\phi_{h(opt)}\right];\bar{Y}_h\mid n_h^*\right)\approx\\&\left(n_h^*\rho\left[1+(n_h^*-1)\rho\right]^{-1}\right)^2(n_h^*)^{-1}\sigma^2(1-\rho)+\left((1-\rho)\left[1+(n_h^*-1)\rho\right]^{-1}\right)^2\sigma^2\rho\\&=\sigma^2\rho(1-\rho)\left[1+(n_h^*-1)\rho\right]^{-1}\end{aligned} \quad (5)$$

3. Optimizing the design

3.1. Area-only optimal design

Provision of precise survey estimates for domains of interest requires that samples of adequate sizes be allocated to the domains. Conflicts arise when equal precision is desired for domains with widely varying population sizes. If estimates of means are desired at the same level of precision for all domains, then an equal allocation may be the most efficient strategy. However, such an allocation can cause a serious loss of efficiency for national estimates.

One way of measuring the performance of designs for small area estimation is with a linear combination of the anticipated MSE's of the small area mean and overall mean estimates. Following Longford (2006), the weights (called inferential priorities) N_h^q for $0 \le q \le 2$ are used. The approximate weighted total of the anticipated MSE's for the areas is given by

$$F = \sum_{h \in U^1} N_h^q E_\xi MSE_\xi \left(\bar{y}_h^C \left[\phi_{h(opt)} \right]; \bar{Y}_h \mid n_h^* \right) + G N_+^{(q)} E_\xi v \qquad (6)$$

where $N_+^{(q)} = \sum_{h \in U^1} N_h^q$

The quantity G is a relative priority coefficient. Ignoring the goal of national estimation corresponds to G = 0 and ignoring the goal of small area estimation corresponds to large values of G, in that case the second component in (6) is dominant. If G is positive, the priority coefficient has to be large because v would generally be much smaller than v_h, where v_h is the sampling variance of the direct estimator $\bar{y}_h$, v is the sampling variance of the synthetic estimator $\bar{y}$, so that G has to be large if the last term of (6) is to have any influence on the outcome. The factor $N_+^{(q)}$ is introduced to appropriately scale for the effect of the absolute sizes of N_h^q and the number of areas on the relative priority G. Criterion (6) is similar to the criterion in Longford (2006), however unlike this paper we adopt the model-assisted approach which treats the design-based inference as the real goal of survey sampling, but employs models to help choose between valid randomization-based alternatives (Särndal et al., 1992). The minimization is subject to a fixed sample size constraint. It would be straightforward to extend this to a fixed cost constraint with cost coefficients specific to the strata.

When national estimation has no priority (G = 0), the solution for the number of units to be sampled from each strata is found by optimizing (6) subject to a fixed total sampling cost function. The stationary point for this optimization is

$$n_{h,opt} = \frac{n\sqrt{N_h^q}}{\sum_{h \in U^1} \sqrt{N_h^q}} + \frac{1-\rho}{\rho} \left(\frac{\sqrt{N_h^q}}{H^{-1} \sum_{h \in U^1} \sqrt{N_h^q}} - 1 \right) \qquad (7)$$

Equation (7) is the optimal design if it gives a feasible solution ($0 \le n_{h,opt} \le N_h$ for all h); if not, the optimal design must be obtained numerically. An approximate solution can be found by setting the non-feasible solutions to $n_{h,opt} = 0$ when $n_{h,opt} < 0$ or $n_{h,opt} = N_h$ when $n_{h,opt} > N_h$ and then reallocating the remaining small areas (Longford, 2006).

In practice it will always be appropriate to set $0 \le q \le 2$, with q = 0 corresponding to all areas being equally important regardless of size, and q = 2 being the best choice for national estimation. In many cases q = 1 would be a sensible compromise.

The first term in (7) above is the optimal allocation for the direct estimator and corresponds to power allocation (Bankier, 1988). The second term will be positive for more populous areas (large N_h) and negative for less populous areas. Therefore, the allocation optimal for composite estimation has more dispersed subsample sizes $n_{h,opt}$ than the allocation that is optimal for direct estimators.

3.2. Compromise design

To incorporate priority for national estimation in optimizing design for small area estimation, we set the relative priority G to a positive value. Unfortunately, this optimization has no simple closed form solution (Molefe, 2012). The solution can be expressed as a quartic equation. Analytic solutions can be found to quartic equations but finding the solution would be convoluted and difficult to interpret. Also, there are up to 4 real-valued solutions. Another approach would be to find a Taylor series approximation based on ρ close to 0 and then minimize this with respect to n_h. The optimal n_h depends on ρ; one could consider n_h to be a function of this quantity and write $n_h = n_h(\rho)$.

The approximate weighted total of the anticipated MSE's for the areas is given by

$$F = \sum_{h \in U^1} N_h^q \sigma^2 \rho (1-\rho)\left[1+(n_h-1)\rho\right]^{-1} + G N_+^{(q)} \sigma^2 \rho (1-\rho) \sum_{h \in U^1} P_h^2 n_h^{-1} \quad (8)$$

Replacing σ^2 by 1 as this value does not affect the optimal design, the minimum of (8) when G>0 satisfies the condition

$$N_h^q \rho^2 \left[1+(n_h-1)\rho\right]^{-2} + G N_+^{(q)} P_h^2 n_h^{-2} = \lambda \quad (9)$$

where λ is the Lagrange multiplier.

This needs to be solved with respect to n_h, but there is no simple closed form solution. One approach would be to find a Taylor series approximation based on

ρ close to 0 to the LHS of (9) and then minimize this with respect to n_h. The objective function is

$$F = \sum_{h\in U^1} N_h^q \rho(1-\rho)\left[1+(n_h-1)\rho\right]^{-1} + GN_+^{(q)}(1-\rho)\sum_{h\in U^1} P_h^2 n_h^{-1} \tag{10}$$

The first derivative with respect to ρ is:

$$F'(\rho) = \sum_{h\in U^1} N_h^q \left\{(1-\rho)\left[1+(n_h-1)\rho\right]^{-1} - \rho\left[1+(n_h-1)\rho\right]^{-1} \right.$$
$$\left. -\rho(1-\rho)n_h\left[1+(n_h-1)\rho\right]^{-2}\right\} - GN_+^{(q)}\sum_{h\in U^1} P_h^2 n_h^{-1}$$

Evaluated at $\rho = 0$, we get:

$$F(0) = GN_+^{(q)}\sum_{h\in U^1} P_h^2 n_h^{-1}$$

$$F'(0) = \sum_{h\in U^1} N_h^q - GN_+^{(q)}\sum_{h\in U^1} P_h^2 n_h^{-1}$$

Hence, the first order Taylor series approximation around $\rho = 0$ is:

$$F(\rho) \approx F(0) + F'(0)\rho$$
$$= GN_+^{(q)}\sum_{h\in U^1} P_h^2 n_h^{-1} + \rho\sum_{h\in U^1} N_h^q - GN_+^{(q)}\rho\sum_{h\in U^1} P_h^2 n_h^{-1}$$
$$= \{(1-\rho)GN_+^{(q)} + \rho\}\sum_{h\in U^1} P_h^2 n_h^{-1} + \rho\sum_{h\in U^1} N_h^q$$

But it is clear that minimizing this approximation of $F(\rho)$ is equivalent to minimizing $\sum_{h\in U^1} P_h^2 n_h^{-1}$, which is equivalent to just ignoring the first term of (10). Therefore, this gives no priority to small area estimation. Hence, a first order Taylor series approximation of F with respect to ρ is not sufficient approximation for the purpose of designing sample for both small areas and national mean.

The second order Taylor series approximation is:

$$F''(\rho) = \sum_{h\in U^1} N_h^q \left\{ -\frac{1}{\left[1+(n_h-1)\rho\right]} - \frac{(1-\rho)n_h}{\left[1+(n_h-1)\rho\right]^2} - \frac{1}{\left[1+(n_h-1)\rho\right]} + \right.$$
$$\left. \frac{\rho n_h}{\left[1+(n_h-1)\rho\right]^2} - \frac{(1-\rho)n_h}{\left[1+(n_h-1)\rho\right]^2} + \frac{\rho n_h}{\left[1+(n_h-1)\rho\right]^2} + \frac{2\rho(1-\rho)n_h^2}{\left[1+(n_h-1)\rho\right]^3} \right\}$$

Evaluated at $\rho = 0$ we get:

$$F''(0) = \sum_{h \in U^1} N_h^q \left\{-1 - n_h - 1 - n_h\right\} = -2\sum_{h \in U^1} N_h^q (n_h + 1)$$

The second order Taylor series approximation is then expressed as:

$$F(\rho) \approx F(0) + F'(0)\rho + \frac{1}{2}F''(0)\rho^2$$
$$= \{(1-\rho)GN_+^{(q)} + \rho\}\sum_{h \in U^1} P_h^2 n_h^{-1} + \rho\sum_{h \in U^1} N_h^q - \rho^2 \sum_{h \in U^1} N_h^q (n_h + 1)$$

We now consider minimizing the second order Taylor series approximation with respect to n_h subject to the cost constraint. The Lagrangian is:

$$L = \{(1-\rho)GN_+^{(q)} + \rho\}\sum_{h \in U^1} P_h^2 n_h^{-1} + \rho\sum_{h \in U^1} N_h^q - \rho^2 \sum_{h \in U^1} N_h^q (n_h + 1) + \lambda(\sum_{h \in U^1} n_h - n)$$

To obtain the solution for the optimal within-strata sample size, we use partial derivatives with respect to n_h and λ, respectively. These are given by equations (A1) and (A2) in the Appendix. The solution for the optimum within-strata sample size n_h is given by

$$\begin{aligned} n_h &\approx n_h(0) + \rho n_h'(0) + \frac{1}{2}\rho^2 n_h''(0) \\ &= nP_h + \frac{1}{2}\rho^2 n^3 P_h (GN_+^{(q)})^{-1}\left\{N_h^q - N^{-1}\sum_{h \in U^1} N_h^{q+1}\right\} \\ &= nP_h\left(1 + \frac{1}{2}\rho^2 n^2 (GN_+^{(q)})^{-1}\left\{N_h^q - N^{-1}\sum_{h \in U^1} N_h^{q+1}\right\}\right) \end{aligned} \tag{11}$$

The approximate solution is a function of G, ρ and q. When $G \uparrow \infty$ the approximate solution for n_h tends to $n_h \approx nP_h$, which is proportional allocation. When G is large, priority is given to estimation of the national mean, hence this is as would be expected, since proportional allocation will be optimal when the focus is on estimating accurately the overall mean. When G = 0 the approximate solution is not defined since division by zero is undefined. The approximate solution is therefore not suitable or appropriate when the only goal is small area estimation. When $\rho \downarrow 0$ the approximate solution is approximately equal to $n_h \approx nP_h$. When $\rho \approx 0$, units within a small area are less similar to each other for the variable of interest. When this happens it is natural for small areas to be represented in proportion to their population sizes.

When q = 1 or 2, it is not clear what the value of the approximate solution will be. The value of n_h depends on the magnitude and sign of $N_h^q - N^{-1}\sum_{h\in U^1} N_h^{q+1}$. We obtain large positive and negative values of n_h depending on the population size of the stratum. For relatively smaller strata, the result is large negative values which would in practice be truncated at zero and the opposite is true for relatively large strata. In practice, these would be truncated to either 0 or the population size.

The approximate analytical optimal design based on $\rho \approx 0$ gave counter-intuitive results, particularly when G is small or zero. Hence, we are now going to approximate n_h based on a different quantity based on both ρ and G rather than on ρ only, say, $n_h = n_h(\alpha)$ where $\alpha = f(\rho, G) = \rho(GN_+^{(q)})^{-1}N^q$. Our interest is the case where α is small. The problem is to minimize

$$F = \sum_{h\in U^1} N_h^q \rho\left[1+(n_h-1)\rho\right]^{-1} + GN_+^{(q)}\sum_{h\in U^1} P_h^2 n_h^{-1}$$

with respect to n_h subject to $\sum_{h\in U^1} n_h = n$. This is equivalent to minimizing

$$F = \alpha\sum_{h\in U^1} P_h^q\left[1+(n_h-1)\rho\right]^{-1} + \sum_{h\in U^1} P_h^2 n_h^{-1}$$

The corresponding Lagrangian function is

$$L = \sum_{h\in U^1} \alpha P_h^q\left[1+\left\{(n_h(\alpha)-1)\rho\right\}\right]^{-1} + \sum_{h\in U^1} P_h^2 n_h^{-1} + \lambda\left(\sum_{h\in U^1} n_h(\alpha) - n\right) \tag{12}$$

The partial derivatives of equation (12) with respect to n_h and λ are, respectively,

$$0 = L_1 = \frac{\partial L}{\partial n_h} = -\alpha P_h^q \rho\left[1+\left\{(n_h(\alpha)-1)\rho\right\}\right]^{-2} - P_h^2 n_h^{-2}(\alpha) + \lambda \tag{13}$$

$$0 = L_2 = \frac{\partial L}{\partial \lambda} = \sum_{h\in U^1} n_h(\alpha) - n \tag{14}$$

Equations (13) and (14) are easily solved when $\lambda = 0$, or in the limit as λ approaches 0. We will derive an approximation for the solution n_h when $\lambda \approx 0$, as this may often be the case in practice.

Let $n_h(\alpha)$ be the solution of (13) and (14) for any given value of α. We can then approximate n_h by $n_h \approx n_h(0) + n_h'(0)\alpha$.

We use equation (13) to obtain the value of $n_h(0)$ by substituting for $\alpha = 0$ to obtain

$$P_h^2 n_h^{-2}(0) = \lambda(0)$$

Solving for $n_h(0)$ we get

$$n_h(0) = P_h\left(\lambda(0)\right)^{-\frac{1}{2}} \tag{15}$$

We substitute for $n_h(0)$ into equation (14) and make it equal to zero to get

$$\left(\lambda(0)\right)^{-\frac{1}{2}} \sum_{h \in U^1} P_h = n = \left(\lambda(0)\right)^{-\frac{1}{2}}$$

Substituting for $\left(\lambda(0)\right)^{-\frac{1}{2}}$ into (15) we obtain the value of $n_h(0)$ as

$$n_h(0) = nP_h \tag{16}$$

We take the derivative of (13) with respect to α :

$$0 = \frac{dL_1}{d\alpha} = \frac{\partial L_1}{\partial \alpha} + \frac{\partial L_1}{\partial n_h}\left(\frac{d}{d\rho} n_h(\alpha)\right) + \frac{\partial L_1}{\partial \lambda}\left(\frac{d}{d\rho}\lambda(\alpha)\right)$$

Therefore

$$\begin{aligned} 0 = \frac{dL_1}{d\alpha} &= -P_h^2 \rho\left[1 + \left\{(n_h(\alpha) - 1)\rho\right\}\right]^{-2} + \\ &\left\{2\alpha P_h^q \rho^2\left[1 + \left\{(n_h(\alpha) - 1)\rho\right\}\right]^{-3} + 2P_h^2 n_h^{-3}(\alpha)\right\} n_h'(\alpha) + \lambda'(\alpha) \end{aligned} \tag{17}$$

where $n_h'(\alpha) = \frac{d}{d\alpha} n_h(\alpha)$ and $\lambda'(\alpha) = \frac{d}{d\alpha}\lambda(\alpha)$. Evaluating (17) at $\alpha = 0$ gives

$$0 = -P_h^q \rho\left[1 + \left\{(n_h(0) - 1)\rho\right\}\right]^{-2} + 2P_h^2 n_h^{-3}(0) n_h'(0) + \lambda'(0)$$

Substituting for $n_h(0)$ given by (16) we get

$$0 = -P_h^q \rho\left[1 + \left\{(nP_h - 1)\rho\right\}\right]^{-2} + 2P_h^{-1} n_h^{-3} n_h'(0) + \lambda'(0)$$

and therefore

$$n_h'(0) = \frac{1}{2} P_h n^3 \left(P_h^q \rho \left[1 + (nP_h - 1)\rho \right]^{-2} - \lambda'(0) \right) \tag{18}$$

Differentiating (14) with respect to α tells us

$$\frac{dL_2}{d\alpha}\Big|_{\alpha=0} = \sum_{h \in U^1} n_h'(0) = 0$$

Combined with (18), this implies that

$$\sum_{h \in U^1} n_h'(0) = \frac{1}{2} P_h n^3 \sum_{h \in U^1} P_h \left(P_h^q \left[1 + (nP_h - 1)\rho \right]^{-2} - \lambda'(0) \right) = 0$$

And therefore $\lambda'(0) = \sum_{h \in U^1} P_h^{q+1} \rho \left[1 + (nP_h - 1)\rho \right]^{-2}$

Substituting for $\lambda'(0)$ into (18) gives

$$n_h'(0) = \frac{1}{2} P_h n^3 \rho \left(P_h^2 \left[1 + (nP_h - 1)\rho \right]^{-2} - \sum_{h \in U^1} P_h^{q+1} \rho \left[1 + (nP_h - 1)\rho \right]^{-2} \right)$$

Hence, the approximation to n_h is

$$n_h \approx n_h(0) + n_h'(0)\alpha$$

$$= nP_h + \frac{1}{2} \alpha P_h n^3 \rho \left(P_h^q \left[1 + (nP_h - 1)\rho \right]^{-2} - \sum_{h \in U^1} P_h^{q+1} \rho \left[1 + (nP_h - 1)\rho \right]^{-2} \right)$$

Substituting for $\alpha = \rho (GN_+^{(q)})^{-1} N^q$ we obtain the general result:

$$n_h \approx nP_h + \frac{1}{2_h} \rho^2 (GN_+^{(q)})^{-1} N^q P_h n^3 \left(P_h^q \left[1 + (nP_h - 1)\rho \right]^{-2} - \sum_{h \in U^1} P_h^{q+1} \left[1 + (nP_h - 1)\rho \right]^{-2} \right)$$

Rewritten, this becomes

$$n_h = nP_h \left(1 + \frac{1}{2} \rho^2 n^2 (GN_+^{(q)})^{-1} N^q \left\{ P_h^q \left[1 + (nP_h - 1)\rho \right]^{-2} - \sum_{h \in U^1} P_h^{q+1} \rho \left[1 + (nP_h - 1)\rho \right]^{-2} \right\} \right) \tag{19}$$

In the previous approximation based on ρ, we obtained large positive or negative values of n_h when n was large. Here, as $n \uparrow \infty$ the approximate sample size is equal to:

$$n_h \approx nP_h\left(1+\frac{1}{2}\rho^2 n^2 (GN_+^{(q)})^{-1} N^q \left\{P_h^q (nP_h\rho)^{-2} - \sum_{h\in U^1} P_h^{q+1}(nP_h\rho)^{-2}\right\}\right)$$

$$= nP_h\left(1+\frac{1}{2}(GN_+^{(q)})^{-1} N^q \left\{P_h^{q-2} - \sum_{h\in U^1} P_h^{q-1}\right\}\right)$$

which seems more reasonable.

When q = 0 and n is large, we get

$$n_h \approx nP_h\left(1+\frac{1}{2}(GH)^{-1}\left\{P_h^{-2} - \sum_{h\in U^1} P_h^{-1}\right\}\right)$$

where $H = N_+^{(0)} = \sum_{h\in U^1} N_h^0$

When q = 1 and n is large, we get

$$n_h \approx nP_h\left(1+\frac{1}{2G}\left\{P_h^{-1} - \sum_{h\in U^1} P_h^0\right\}\right)$$

When q = 2 and n is large, we get

$$n_h \approx nP_h\left(1+\frac{1}{2}(GN_+^{(2)})^{-1} N^2 \left\{P_h^0 - \sum_{h\in U^1} P_h^1\right\}\right) = nP_h$$

A priority exponent of q = 2 implies proportional allocation, hence the result is as expected.

When $G \uparrow \infty$ the approximate sample size is equal to $n_h \approx nP_h$. This result is as expected since very large G implies more priority for national estimation. Proportional allocation will be optimal when the focus is on estimating accurately the overall mean. When $G \downarrow 0$, this corresponds to $\alpha \uparrow \infty$, and the approximate solution is undefined. This means that the alternative approximate analytical optimal design for n_h breaks down as $G \downarrow 0$. Perhaps this is not surprising, as our approximation is based on small α not large α. When $\rho \downarrow 0$ the alternative

approximate analytical design is equal to $n_h \approx nP_h$. When $\rho \approx 0$, units within a small area are somewhat similar to each other though the degree of similarity is very low. Hence, it is appropriate for sample sizes within small areas to be in proportion to their population sizes.

4. Numerical example

We use data on the 26 cantons of Switzerland (Longford, 2006); their population sizes range from 15,000 (Appenzell-Innerrhoden) to 1.23 million (Zurich). The population of Switzerland is 7.26 million. We assume that n = 10,000, ρ = 0.025. We allocate a sample to the 26 cantons in Switzerland for q = 1 and a range of values of $G \in \{50, 100, 200, 500\}$ using the approximation in equation (11). The planned overall sample size is n = 10,000. The result of the percentiles of the sample sizes is shown in Table 4.1.

Table 4.1. Canton sample sizes by Taylor approximation when q = 1 and $\rho = 0.025$

	Percentiles of n_h				
Priority Coefficient	**Minimum**	**1st Quarter**	**Median**	**3rd Quarter**	**Maximum**
G = 50	-9322.0	-8620.0	-5648.0	-2226.0	97380.0
G = 100	-4470.0	-4097.0	-2634.0	-1088.0	49540.0
G = 200	-2050.0	-1878.0	-1126.0	-519.2	25620.0
G = 500	-617.0	-584.2	-324.5	-168.2	11260.0

When G > 0 and q = 1, the solution gives negative sample sizes for smaller cantons and very large positive sample size for the largest canton so that the negative sample sizes will be truncated at zero.

In summary, the approximate analytical optimal design based on $\rho \approx 0$ does not seem like a sensible approximation as evidenced by the allocation in Table 4.1.

We similarly allocate the sample sizes using the approximation in equation (19). The planned overall sample size is n = 10,000. The result of the percentiles of the sample sizes is shown in Table 4.2.

Table 4.2. Canton sample sizes by the alternative Taylor approximation when q = 1 and $\rho = 0.025$

	Percentiles of n_h				
Priority Coefficient	**Minimum**	**1st Quarter**	**Median**	**3rd Quarter**	**Maximum**
G = 50	237.0	275.2	296.0	383.8	1152.0
G = 100	129.0	181.5	290.5	426.8	1422.0
G = 200	75.0	139.0	288.0	448.2	1558.0
G = 500	42.0	113.2	286.5	461.2	1639.0

From Table 4.2, we see that when G = 50 the sample sizes of the least populous cantons are boosted in relation to proportional allocation at the expense of the most populous cantons. As G increases the sample size allocation approaches proportional allocation.

In summary, the alternative approximate analytical design seems to be useful especially when G > 0. The design seems sensible when there is priority for national estimation and is not applicable when the only priority is small area estimation.

5. Conclusions

The anticipated MSE is a sensible objective criterion for sample design because the particular sample which will be selected is not available in advance of the survey. Hence, a criterion which averages over all possible samples is appropriate. Särndal et al. (1992, Chapter 14) base their optimal designs on the anticipated variance, which similarly averages over both model realizations and sample selection, although they consider only approximately design-unbiased estimators.

An analytical solution for the stationary point exists when the only priority is small area estimation. However, there are difficulties in applying it because when the strata have disparate population sizes, the stationary point gives negative sample sizes so that the optimum must be obtained numerically. The numerical optimum then has some strata with $n_h = 0$ which is also not desirable.

When priority is given to national estimation as well as to small area estimation so that G > 0, two approximate solutions were derived, based on $\rho \approx 0$ and $\alpha = f(\rho, G) = \rho (G N_+^q)^{-1} N^q \approx 0$. Both had undesirable properties, giving very large positive and negative sample sizes in some cases. This approximate solution gives counter-intuitive results, with large negative or positive values when there are unequal priorities for strata. Therefore, the Taylor approximation is not useful. An undesirable property of the second design is that it is not applicable when there is no priority for national estimation (G = 0).

REFERENCES

BANKIER, M. D., (1988). Power Allocations: Determining Sample Sizes for Sub-national Areas. The American Statistician, 42(3):174−177.

BINMORE, K. G., (1982). Mathematical Analysis: A straightforward approach. Cambridge University Press, 2nd edition.

BROCK, D. B., FRENCH, D. K., PEYTON, B. W., (1980). Small Area Estimation: Empirical Evaluation of Several Estimators for Primary Sampling Units. In Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 766−771.

DEMIDOVICH, B., editor, (1964). Problems in Mathematical Analysis. MIR Publishers.

ERICKSEN, E. P., (1973). Recent Developments in Estimation for Local Areas. In Proceedings of the Section on Social Statistics, American Statistical Association, pp. 37−41.

FULLER, W. A., (1999). Environmental Surveys Over Time. Journal of Agricultural, Biological and Environmental Statistics, 4: 331−345.

GONZALEZ, M. E., (1973). Use and Evaluation of Synthetic Estimates. In Proceedings of the Section on Social Statistics, American Statistical Association, pp. 33−36.

HIDIROGLOU, M. A., PATAK, Z., (2004). Domain estimation using linear regression. Survey Methodology, 30: 67−78.

LONGFORD, N. T. (2006). Sample Size Calculation for Small-Area Estimation. Survey Methodology, 32(1): 87−96.

MOLEFE, W. B., (2012). Sample Design for Small Area Estimation. PhD thesis, University of Wollongong, http://ro.uow.edu.au/theses/3495.

MOLEFE, W. B., CLARK, R. G., (2015). Model-Assisted Optimal Allocation For Planned Domains Using Composite Estimation, Survey Methodology (forthcoming).

RAO, J. N. K., (2003). Small Area Estimation. Wiley.

ROYALL, R. M., (1973). Discussion of two Papers on Recent Developments in Estimation of Local Areas. In Proceedings of the Section on Survey Research Methods, American, Statistical Association, pp. 43−44.

SĂRNDAL, C., SWENSSON, B., WRETMAN, J., (1992). Model Assisted Survey Sampling. Springer-Verlag.

SCHAIBLE, W. L., (1978). Choosing Weight for Composite Estimators for Small Area Statistics. In Proceedings of the Section on Survey Research Methods, American Statistical 3 Association, pp. 741−746.

SINGH, M. P., GAMBINO, J., MANTEL, H. J., (1994). Issues and Strategies for Small Area Data. Survey Methodology, 20(1): 3−22.

APPENDIX

$$0 = L_1 = \frac{\partial L}{\partial n_h} = -\{(1-\rho)+\rho\}GN_{+}^{(q)}P_h^2 n_h^{-2} - \rho^2 N_h^q + \lambda \tag{A1}$$

$$0 = L_1 = \frac{\partial L}{\partial \lambda} = \sum\nolimits_{h\in U^1} n_h - n \tag{A2}$$

Equations (A1) and (A2) are easily solved when $\rho = 0$, or in the limit as ρ approaches 0. We will derive an approximation for the solution n_h when $\rho \approx 0$, as this may often be the case in practice.

Let $n_h(\rho)$ be the solution of (A1) and (A2) for any given value of ρ. We can then approximate n_h by $n_h \approx n_h(0) + n_h'(0)\rho + \frac{1}{2}n_h''(0)\rho^2$

It is easily shown that $n_h(0) = nP_h$. To derive $n_h(0)$ we use (A1) to obtain the value of $n_h(0)$ by substituting for $\rho = 0$ to obtain $GN_{+}^{(q)}P_h^2 n_h^{-2}(0) = \lambda(0)$

Solving for $n_h(0)$ we get

$$n_h(0) = P_h\left(\frac{GN_{+}^{(q)}}{\lambda(0)}\right)^{\frac{1}{2}} \tag{A3}$$

Substituting for n_h into (A2) gives $\left(\frac{GN_{+}^{(q)}}{\lambda(0)}\right)^{\frac{1}{2}}\sum_{h\in U^1} P_h = n$. Summing through and re-arranging terms gives $\left(\lambda(0)\right)^{-\frac{1}{2}} = n(GN_{+}^{(q)})^{-\frac{1}{2}}$. Substituting for $\left(\lambda(0)\right)^{-\frac{1}{2}}$ into (A3), we obtain the value of $n_h(0)$ as

$$n_h(0) = nP_h \tag{A4}$$

We take the first derivative of (A1) with respect to ρ:

$$0 = \frac{dL_1}{d\rho} = \frac{\partial L_1}{\partial \rho} + \frac{\partial L_1}{\partial n_h}\left(\frac{d}{d\rho}n_h(\rho)\right) + \frac{\partial L_1}{\partial \lambda}\left(\frac{d}{d\rho}\lambda(\rho)\right) \tag{A5}$$

using the result on differentiation of composite functions by, for example, Demidovich (1964) and Binmore (1982). The partial derivative of (A1) with respect to ρ is:

$$\frac{\partial L_1}{\partial \rho} = GN_+^{(q)} P_h^2 n_h^{-2} - 2\rho N_h^q$$

and the partial derivative with respect to n_h gives

$$\frac{\partial L_1}{\partial n_h} = 2(1-\rho) GN_+^{(q)} P_h^2 n_h^{-3}$$

Substituting the partial derivatives in (A5) gives

$$\begin{aligned} 0 = \frac{dL_1}{d\rho} &= \frac{\partial L_1}{\partial \rho} + \frac{\partial L_1}{\partial n_h}\left(\frac{d}{d\rho} n_h\right) + \frac{\partial L_1}{\partial \lambda}\left(\frac{d}{d\rho}\lambda\right) \\ &= GN_+^{(q)} P_h^2 n_h^{-2} - 2\rho N_h^q + 2(1-\rho) GN_+^{(q)} P_h^2 n_h^{-3} n_h'(\rho) + \lambda'(\rho) \end{aligned} \tag{A6}$$

where $n_h'(\rho) = \frac{d}{d\rho} n_h$ and $\lambda'(\rho) = \frac{d}{d\rho}\lambda$.

Evaluating (A6) at $\rho = 0$ gives

$$0 = GN_+^{(q)} P_h^2 n_h^{-2}(0) + 2GN_+^{(q)} P_h^2 n_h^{-3}(0) n_h'(0) + \lambda'(0) \tag{A7}$$

Solving for $n_h'(0)$ gives:

$$\begin{aligned} n_h'(0) &= -2\left\{\lambda'(0) + GN_+^{(q)} P_h^2 n_h^{-2}(0)\right\}(GN_+^{(q)} P_h^2)^{-1} n_h^{-3}(0) \\ &= -2\lambda'(0)(GN_+^{(q)} P_h^2)^{-1} n_h^{-3}(0) - 2GN_+^{(q)} P_h^2 n_h^{-2}(0)(GN_+^{(q)} P_h^2)^{-1} n_h^{-3}(0) \\ &= -\lambda'(0)(GN_+^{(q)})^{-1} n^3 P_h - nP_h \end{aligned} \tag{A8}$$

Differentiating (A2) with respect to ρ tells us

$$\left.\frac{dL_2}{d\rho}\right|_{\rho=0} = \sum\nolimits_{h \in U^!} n_h'(0) = 0 \tag{A9}$$

Combined with (A8), this implies that

$$\sum_{h\in U^1}\left\{-\lambda'(0)(GN_+^{(q)})^{-1}n^3P_h - nP_h\right\} = 0$$

Consequently, $\lambda'(0) = -GN_+^{(q)}n^{-2}$. Substituting for $\lambda'(0)$ into (A8) gives the result that

$$n_h(0) = 0$$

We now take the second derivative of (A5). Let

$$L_3 = \frac{dL_1}{d\rho} = \frac{\partial L_1}{\partial \rho} + \frac{\partial L_1}{\partial n_h}\left(\frac{d}{d\rho}n_h(\rho)\right) + \frac{\partial L_1}{\partial \lambda}\left(\frac{d}{d\rho}\lambda(\rho)\right)$$

and therefore

$$L_3 = GN_+^{(q)}P_h^2n_h^{-2} - 2\rho N_h^q + 2(1-\rho)GN_+^{(q)}P_h^2n_h^{-3}n_h'(\rho) + \lambda'(\rho) \qquad \text{(A10)}$$

We take the derivative of (A10) with respect to ρ:

$$0 = \frac{dL_3}{d\rho} = \frac{\partial L_3}{\partial \rho} + \frac{\partial L_3}{\partial n_h}\left(\frac{d}{d\rho}n_h'(\rho)\right) + \frac{\partial L_3}{\partial \lambda}\left(\frac{d}{d\rho}\lambda'(\rho)\right)$$

The partial derivative of (A10) with respect to ρ is given by

$$0 = \frac{dL_3}{d\rho} = -2N_h^q - 2GN_+^{(q)}P_h^2n_h^{-3}n_h'(\rho) + 2(1-\rho)GN_+^{(q)}P_h^2n_h^{-3}\frac{\partial}{\partial\rho}n_h'(\rho) + \frac{\partial}{\partial\rho}\lambda'(\rho)$$

The partial derivative evaluated at $\rho = 0$ is then

$$\frac{\partial L_3}{\partial \rho}|_{\rho=0} = -2N_h^q - 2GN_+^{(q)}P_h^2n_h^{-3}(0)n_h'(0) + 2GN_+^{(q)}P_h^2n_h^{-3}(0)n_h''(0) + \lambda''(0)$$
$$= -2N_h^q + 2GN_+^{(q)}P_h^2(nP_h)^{-3}n_h''(0) + \lambda''(0)$$

since $n_h'(0) = 0$.

The partial derivative of (A10) with respect to n_h is given by

$$\frac{\partial L_3}{\partial n_h} = -2GN_+^{(q)}P_h^2 n_h^{-3} - 6(1-\rho)GN_+^{(q)}P_h^2 n_h^{-4} n_h'(\rho) + 2(1-\rho)GN_+^{(q)}P_h^2 n_h^{-3} n_h''(\rho)$$

The partial derivative evaluated at $\rho = 0$ is:

$$\begin{aligned} 0 &= \frac{\partial L_3}{\partial n_h} \\ &= -2GN_+^{(q)}P_h^2 n_h^{-3}(0) + 2GN_+^{(q)}P_h^2 n_h^{-3}(0) n_h''(0) \\ &= -GN_+^{(q)}P_h^2 (nP_h)^{-3} + GN_+^{(q)}P_h^2 (nP_h)^{-3} n_h''(0) \end{aligned}$$

since $n_h'(0) = 0$.

We put the results together to obtain

$$\begin{aligned} 0 &= \frac{\partial L_3}{\partial \rho}|_{\rho=0} \\ &= \left\{ \frac{\partial L_3}{\partial \rho} + \frac{\partial L_3}{\partial n_h}\left(\frac{d}{d\rho} n_h(\rho) \right) + \frac{\partial L_3}{\partial \lambda}\left(\frac{d}{d\rho} \lambda(\rho) \right) \right\} \\ &= -2N_h^q + 2GN_+^{(q)}P_h^2 n_h^{-3} P_h^{-1} n_h''(0) + \lambda''(0) \end{aligned}$$

since $n_h'(0) = \frac{d}{d\rho} n_h(\rho)|_{\rho=0} = 0$ and $\lambda'(0) = \frac{d}{d\rho}\lambda(\rho)|_{\rho=0} = 0$.

Solving for $n_h''(0)$ we find

$$n_h''(0) = (2N_h^q - \lambda''(0))\frac{1}{2}(GN_+^{(q)})^{-1} n^3 P_h \qquad \text{(A11)}$$

Differentiating (A9) with respect to ρ gives:

$$0 = \frac{d}{d\rho}\left(\frac{dL_2}{d\rho} \right)|_{\rho=0} = \frac{d}{d\rho}\sum_{h\in U^1} n_h'(\rho)|_{\rho=0} = \sum_{h\in U^1} n_h''(0)$$

Combined with (A11), this implies that

$$\sum_{h\in U^1} n_h''(0) = \sum_{h\in U^1} (2N_h^q - \lambda''(0))\frac{1}{2}(GN_+^{(q)})^{-1} n^3 P_h = 0$$

Therefore

$$(GN_+^{(q)})^{-1} n^3 N^{-1} \sum_{h\in U^1} N_h^{q+1} - \frac{1}{2}\lambda''(0)(GN_+^{(q)})^{-1} n^3 \sum_{h\in U^1} P_h = 0$$

Solving for $\lambda''(0)$ we get $\lambda''(0) = 2N^{-1}\sum_{h\in U^1} N_h^{q+1}$. Substituting into (A11) gives

$$n_h''(0) = (N_h^q - N^{-1}\sum_{h\in U^1} N_h^{q+1})(GN_+^{(q)})^{-1} n^3 P_h$$

Hence, our approximation to n_h is:

$$\begin{aligned} n_h &\approx n_h(0) + \rho n_h'(0) + \frac{1}{2}\rho^2 n_h''(0) \\ &= nP_h + \frac{1}{2}\rho^2 n^3 P_h (GN_+^{(q)})^{-1}\left\{N_h^q - N^{-1}\sum_{h\in U^1} N_h^{q+1}\right\} \\ &= nP_h\left(1 + \frac{1}{2}\rho^2 n^2 (GN_+^{(q)})^{-1}\left\{N_h^q - N^{-1}\sum_{h\in U^1} N_h^{q+1}\right\}\right) \end{aligned}$$